

PCT/JP 00/03625

09/762126

02.06.00

JP00/3625

4 日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application:

1999年 7月27日

27 JUL 2000

出 願 番 号

Application Number:

平成11年特許願第212501号

出 願 人

Applicant (s):

セイコーエプソン株式会社

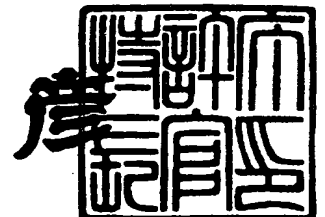
PRIORITY
DOCUMENT

SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

2000年 6月29日

特許庁長官
Commissioner,
Patent Office

近 藤 隆 彦



出証番号 出証特2000-3052064

【書類名】 特許願

【整理番号】 J0074221

【提出日】 平成11年 7月27日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 7/36
G06F 17/30

【発明の名称】 文書分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体

【請求項の数】 15

【発明者】
【住所又は居所】 長野県諏訪市大和3丁目3番5号 セイコーエプソン株式会社内
【氏名】 三輪 真司

【特許出願人】
【識別番号】 000002369
【氏名又は名称】 セイコーエプソン株式会社
【代表者】 安川 英昭

【代理人】
【識別番号】 100093388
【弁理士】
【氏名又は名称】 鈴木 喜三郎
【連絡先】 0266-52-3139

【選任した代理人】
【識別番号】 100095728
【弁理士】
【氏名又は名称】 上柳 雅誉

【選任した代理人】
【識別番号】 100107261
【弁理士】

【氏名又は名称】 須澤 修

【手数料の表示】

【予納台帳番号】 013044

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9711684

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体

【特許請求の範囲】

【請求項 1】 複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも 2 つのクラスタを統合するクラスタマージ処理を行うことを特徴とする文書分類方法。

【請求項 2】 前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基にクラスタ間の関連性を判断してクラスタマージすることを特徴とする請求項 1 記載の文書分類方法。

【請求項 3】 前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタそれぞれを特徴づける特徴要素が、そのクラスタマージ処理対象となるそれぞれのクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージすることを特徴とする請求項 1 記載の文書分類方法。

【請求項 4】 前記クラスタマージ処理は、少なくとも 2 つのクラスタ間で行い、一回目のクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起これなくなるまでそれを繰り返すことを特徴とする請求項 1 から 3 のいずれか 1 項に記載の文書分類方法。

【請求項 5】 前記クラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することを特徴とする請求項 1 から 4 のいずれか 1 項に記載の文書分類方法。

【請求項 6】 複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、

このクラスタリング部により得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部と、

を有することを特徴とする文書分類装置。

【請求項7】 前記クラスタマージ部が行うクラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基にクラスタ間の関連性を判断してクラスタマージすることを特徴とする請求項6記載の文書分類装置。

【請求項8】 前記クラスタマージ部が行うクラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタそれぞれを特徴づける特徴要素が、そのクラスタマージ処理対象となるそれぞれのクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージすることを特徴とする請求項6記載の文書分類装置。

【請求項9】 前記クラスタマージ部が行うクラスタマージ処理は、少なくとも2つのクラスタの組み合わせで行い、一回目のクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すことを特徴とする請求項6から8のいずれか1項に記載の文書分類装置。

【請求項10】 前記クラスタマージ部がクラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することを特徴とする請求項6から9のいずれか1項に記載の文書分類装置。

【請求項11】 複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類処理プログラムを記録した記録媒体であって、その文書分類処理プログラムは、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング処理手順と、

これにより分類された複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理手順と、

を含むことを特徴とする文書分類処理プログラムを記録した記録媒体。

【請求項12】 前記クラスタマージ処理手順で行われるクラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基にクラスタ間の関連性を判断してクラスタマージすることを特徴とする請求項11記載の文書分類処理プログラムを記録した記録媒体。

【請求項13】 前記クラスタマージ処理手順で行われるクラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタそれぞれを特徴づける特徴要素が、そのクラスタマージ処理対象となるそれぞれのクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージすることを特徴とする請求項11記載の文書分類処理プログラムを記録した記録媒体。

【請求項14】 前記クラスタマージ処理手順で行われるクラスタマージ処理は、少なくとも2つのクラスタ間で行い、一回目のクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すことを特徴とする請求項11から13のいずれか1項に記載の文書分類処理プログラムを記録した記録媒体。

【請求項15】 前記クラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することを特徴とする請求項11から14のいずれか1項に記載の文書分類処理プログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は多数の文書を意味的に共通性を有する複数のクラスタに分類する文書

分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体に関する。

【0002】

【従来の技術】

多数の文書を意味的なまとまりごとの複数のクラスタに分類する際、それぞれの文書から特徴要素を抽出し、その特徴要素に基づいて分類することが行われている。その分類手法として、それぞれの文書全体（表題や本文など1つの文書を構成する文書内容全体）を特徴要素の抽出対象とし、それぞれの文書全体から特徴要素を抽出し、抽出された特徴要素に基づいて複数のクラスタに分類する文書分類方法がある。

【0003】

この文書全体を特徴要素抽出の対象として分類を行うと、文書の形態素解析や、特徴抽出処理が非常に繁雑であり、CPUがその処理を行う場合、CPUに対する負荷を大きいものとしている。また、一般に、文書はその文書の主旨とは直接関係のない記述を多く含んでいるのが普通である。したがって、文書全体を特徴要素抽出の対象とすると、それによって分類されたクラスタは情報の分類という観点から見たとき、あまり意味のない分類となることも多い。つまり、ノイズクラスタが多数生成されてしまうということにもなる。

【0004】

このような問題点を解消する手法として、それぞれの文書の主旨を適切に表す部分としてそれぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法がある。この手法は、文書の主旨を適切に反映した文書分類を可能とすることができるものとして期待される。

【0005】

このように、従来から文書を幾つかのクラスタに分類する手法は幾つか考えられている。

【0006】

【発明が解決しようとする課題】

しかしながら、上述した適切な分類がなされる手法である文書の表題部から抽出された特徴要素に基づいて文書を分類する手法を用いたとしても、それによって得られる分類結果は、クラスタの数が多くなりすぎることもあり、ユーザ側から見たときに、決して適切な分類が行われたとは思えない場合もでてくる。たとえば、分類されて得られる多数のクラスタを比較した場合、それぞれのクラスタに共通した文書が数多く含まれる場合もある。このような場合、ユーザは提示された多数のクラスタについて、結局は、自分で整理し、その中から自分の本当に欲しい情報を探すというような面倒な処理を行うことになる。

【0007】

そこで、本発明は、分類結果として得られた多数のクラスタに対してクラスタマージ処理を行うことで、より一層、ユーザにとってわかりやすく簡潔的に分類された分類結果を提示できるようにすることを目的としている。

【0008】

【課題を解決するための手段】

前述の目的を達成するために、本発明の文書分類方法は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行うようにしている。

【0009】

また、本発明の文書分類装置は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、このクラスタリング部により得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部とを有する構成としている。

【0010】

また、本発明の文書分類処理プログラムを記録した記録媒体は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類処理プログラムを記録した記録媒体であって、その文書分類処理プログラムは、前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング処理手順と、これにより分類された複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理手順とを含むものである。

【0011】

これら各発明において、前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基にクラスタ間の関連性を判断してクラスタマージする処理である。

【0012】

また、前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタそれぞれを特徴づける特徴要素が、そのクラスタマージ処理対象となるそれぞれのクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージする処理であってもよい。

【0013】

そして、これらクラスタマージ処理は、少なくとも2つのクラスタ間で行い、一回目のクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すようにする。

【0014】

さらに、前記クラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力する。

【0015】

このように本発明は、それぞれの文書を複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書の内容に基づいてそれ

それぞれのクラスタ間の関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行うようにしている。これによって、最初のクラスタリング処理によって、多数のクラスタが生成されたとしても、それぞれのクラスタ間でクラスタ同志の関連性を判断し、関連性の高い複数のクラスタを統合することができるので、簡潔化された分類結果をユーザに提示することができ、ユーザは自分の欲しい情報を効率よく探すことができるようになる。

【0016】

また、クラスタ間の関連性の判断は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、各々のクラスタに共通して含まれる文書数を基にして行うので、簡単で的確なクラスタマージ処理を行うことができる。

【0017】

また、クラスタ間の関連性の判断を行うための他の方法として、特徴要素がクラスタマージ処理対象となるクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージ処理を行うようにしてもよく、これによれば、実際の文書内容に基づいたクラスタ同志の関連性の判断が行えるので、適切なクラスタマージ結果を得ることができる。

【0018】

そして、クラスタマージ処理は、少なくとも2つのクラスタの組み合わせで行い、さらに、所定の数のクラスタ間でのクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すことによって、最終的には、より簡潔的に整理された分類結果を得ることができる。

【0019】

また、このようなクラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することにより、ユーザはどのような状況でクラスタマージ処理がなされたかを知ることができるので、クラスタマージ処理後の結果から自分の欲しい情報を探す際に、その付加状況を参考にして探すことができる。

【0020】

【発明の実施の形態】

以下、本発明の実施の形態について説明する。なお、この実施の形態で説明する内容は、本発明の文書分類方法および文書分類装置についての説明であるとともに、本発明の文書分類処理プログラムを記録した記録媒体における文書分類処理プログラムの具体的な処理内容をも含むものである。

【0021】

また、この実施の形態では、文書分類の手法として、前述したように、それぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法を用いるものとする。

【0022】

図1は本発明を実現するための装置構成を示すもので、大きく分けると、複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部1と、このクラスタリング部1により得られた複数のクラスタ間で各々のクラスタに含まれる文書の内容に基づいて各々のクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部2と、このクラスタマージ部2でクラスタマージ処理された分類結果を出力する分類結果出力部3とを有した構成となっている。

【0023】

クラスタリング部1は、文書記憶部11、文解析部12、特徴要素抽出部13、特徴テーブル作成部14、文書分類部15、分類結果記憶部16を有している。

【0024】

クラスタマージ部2はクラスタを統合するものであるがこれについての処理内容については後に詳細に説明する。

【0025】

分類結果出力部3は、出力制御部31、表示部32を有し、クラスタマージ部2によるクラスタマージ処理結果を出力させるための制御を行う。

【0026】

上述のクラスタリング部1に含まれる文書記憶部11はこの場合、多数の文書データをデータベースとして持つものである。ここでは、たとえば、図2に示すような文書群を分類する場合を説明する。図2に示される文書群は、それぞれが独立した文書D1, D2, ..., D7を有し、これらの文書D1, D2, ..., D7は表題部T1, T2, ..., T7と、それに対する本文A1, A2, ..., A7を持っているものとする。

【0027】

文解析部12は文書記憶部11に記憶されている文書を文解析し、それぞれの文書の表題部を検出する。この文解析部12が行う表題部の検出は、具体的には次のようにして行う。

【0028】

まず、第1の方法として、文書構造様式によって表題と規定される部分があればその部分を表題部とする。また、第2の方法として、文書構造様式によって、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とする。また、第3の方法として、定められた数の文または単語を文書先頭より抽出し、その抽出した部分を表題部とする。さらには、これら第1、第2、第3の方法を順次行い、第1の方法を行ったとき、表題と規定されている部分があればその部分を表題部とし、表題と規定される部分が存在しなければ、第2の方法を行い、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とし、標準より大きな文字で表示する指定がなされていなければ、第3の方法を行って表題部を検出する。

【0029】

特徴要素抽出部13は、文解析部2で検出されたそれぞれの文書の表題部の中から特徴要素を抽出する。

【0030】

特徴テーブル作成手段14は、前記表題部から抽出された特徴要素とそれぞれの文書との関係を示す特徴テーブルを作成する。なお、この特徴テーブルの具体的な内容については後述する。

【0031】

文書分類部15は、前述の特徴テーブルの内容を参照し、文書D1, D2, . . . , D7を意味的に共通性のある複数のクラスタに分類する。つまり、文書D1, D2, . . . , D7の表題部に存在する特徴要素に基づいて、共通する特徴要素を持つ処理対象文書を1つのまとまりとし、そのまとまりを1つのクラスタとする。なお、この文書分類部15は同義特徴辞書（図示せず）を有し、共通する特徴要素を持つ処理対象文書を1つのまとまりとする処理を行う際、共通する特徴要素であるか否かの判断を、その同義語辞書を用い同義語が有るか否かにより行い、同義語が存在する場合にはそれを同じクラスタとする処理を行うことも可能である。

【0032】

分類結果記憶部16は、文書分類部15によって分類された内容を記憶する。

【0033】

このような構成において、本発明の文書分類処理について説明する。本発明が行う概略的な文書分類処理は、図3のフローチャートに示すように、処理対象となる多数の文書を意味的に共通性を有する複数のクラスタに分類し（ステップS1）、これにより分類された複数のクラスタ間で各々のクラスタに含まれる文書に基づいて（これについては後に説明する）それぞれのクラスタの関連性を判断する（ステップS2）。そして、一定以上の関連性を有する少なくとも2つのクラスタを統合する（ステップS3）。以下、具体例を参照して詳細に説明する。

【0034】

ここでは、図2で示した文書D1, D2, . . . , D7を分類する例について説明する。この実施の形態では、それぞれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理された結果についてクラスタマージ処理を行う。まず始めに、表題部から特徴要素を抽出し、その抽出された特徴要素に基づいて行われるクラスタリング処理（クラスタリング部1が行う処理）について説明する。

【0035】

これらの文書D1, D2, . . . , D7は、文解析部12にて表題部が検出さ

れる。たとえば、文書D1については表題部T1が検出され、文書D2については表題部T2が検出され、文書D3については表題部T3が検出されるというように、それぞれの文書D1, D2, ..., D7の表題部T1, T2, ..., T7が検出される。

【0036】

そして、特徴要素抽出部13によって、それぞれの表題部に存在する特徴要素が抽出されたのち、特徴テーブル作成部14により、それぞれの特徴要素とその特徴要素を表題部に含む文書との関係を示す特徴テーブルが作成される。この特徴テーブルの例を図4に示す。なお、ここでは、文書数が3つ以上取り出される特徴要素とその特徴要素を含む文書との関係を示し、特徴テーブル内に示される数値は、その特徴要素が各文書の表題部に幾つ含まれているかの数を示している。たとえば、「用紙」という特徴要素は、文書D1, D4, D6, D7のそれぞれの表題部に、それぞれ1個ずつ含まれていることを示している。

【0037】

図4の特徴テーブルからもわかるように、表題部に「用紙」という特徴要素を含む文書は、文書D1, D4, D6, D7であり、また、表題部に「カセット」という特徴要素を含む文書は、文書D1, D4, D7であり、さらに、表題部に「増設」という特徴要素を含む文書は、文書D2, D3, D5, D7である。なお、図2において、これら各特徴要素部分にはアンダーラインが施されている。

【0038】

そして、文書分類部15はこのような特徴テーブルを参照して、それぞれの特徴要素ごとの文書クラスタ分けを行う。その分類結果を図5に示す。なお、このようなクラスタに分類する際、前述したように、共通する特徴要素であるか否かの判断を、同義語辞書を用い同義語が有るか否かによっても行い、同義語が存在する場合にはそれを同じ文書クラスタとする処理を行うことも可能である。たとえば、「用紙」と「印刷紙」の両方が特徴要素として抽出されたとすれば、これらの特徴要素を表題部に含む文書は同じクラスタとするなどという処理を行う。

【0039】

このような分類結果は分類結果記憶部16に格納される。図5に示される分類

結果において、たとえば、「用紙」で分類されたクラスタ（文書D1, D4, D6, D7が含まれる）について見れば、図2の文書内容からもわかるように、文書D1は用紙カセットについての内容であり、文書D4は用紙設定についての内容であり、文書D6は印刷された後の用紙の汚れについての内容であり、文書D7は用紙カセットの増設についての内容である。

【0040】

このように、これらの文書D1, D4, D6, D7はどれも用紙に関する内容であり、1つのクラスタとして分類されて何等问题のないものとなり、その分類結果は適切であるといえる。

【0041】

また、「カセット」で分類されたクラスタ（文書D1, D4, D7が含まれる）について見れば、図2の文書内容からもわかるように、文書D1は用紙カセットについての内容であり、文書D4は用紙設定についての内容であり、文書D7は用紙カセットの増設についての内容である。

【0042】

このように、これらの文書D1, D4, D6, D7にはどれも用紙をセットすることに関する内容が含まれており、1つのクラスタとして分類されて何等问题のないものとなり、その分類結果は適切であるといえる。

【0043】

また、「増設」で分類されたクラスタ（文書D2, D3, D5, D7が含まれる）について見れば、図2の文書内容からもわかるように、文書D2はメモリの増設についての内容であり、文書D3はインタフェースカードの増設についての内容であり、文書D5はハードディスクの増設についての内容であり、文書D7は用紙カセットの増設についての内容である。

【0044】

このように、これらの文書D2, D3, D5, D7はどれも何かを増設する場合についての内容であり、1つのクラスタとして分類されて何等问题のないものとなり、その分類結果は適切であるといえる。

【0045】

このような適切な分類が行える理由としては、それぞれの文書の表題部から特徴要素を抽出し、その特徴要素に基づいて文書を分類しているからである。つまり、文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることが多い。したがって、文書の表題部に含まれる特徴要素を用いて分類を行うことにより、分類結果が散漫になることが少なく、また、ノイズクラスタが生成される率も少なくすることができる。また、各文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることから、文書の制作者側の視点による分類が得られる。

【0046】

そして、分類が行われた後、ユーザによって、たとえば、「用紙」についてのクラスタの選択指示が出されたとすると、そのクラスタに属する文書D1, D4, D6, D7が文書記憶部11から読み出されて表示部32に表示される。なお、このときの表示内容としては、前述したように、文書番号や文書名のみでもよく、さらには、その文書内容を表示させるようにしてもよい。

【0047】

ところで、本発明は以上のようにクラスタリング処理した結果について、さらに、クラスタマージ部2によってクラスタマージ処理を行う。

【0048】

すなわち、図5に示す分類結果において、特徴要素である「用紙」と「カセット」について見ると、「用紙」のクラスタには文書D1, D4, D6, D7が含まれ、「カセット」のクラスタには文書D1, D4, D7に存在することがわかる。

【0049】

このように、「用紙」のクラスタと「カセット」のクラスタには、共に文書D1, D4, D7が共通して存在している。これは、「用紙」という特徴要素と「カセット」という特徴要素は相互に関連した状態で用いられることが多いことを意味している。たとえば、文書D1, D4, D7の表題部または本文のなかに「用紙カセット」という用語が用いられている。つまり、これらの文書D1, D4

、D7は共通性の高い文書であり、これら文書D1、D4、D7は同じクラスタに分類した方がより好ましいと考えられる。

【0050】

これを実現するために本発明では特徴要素に基づいてクラスタリングしたあと、そのクラスタリング結果に対しクラスタマージ処理を施す。

【0051】

このクラスタマージ処理について以下に説明する。まず始めに、図5の分類結果とは関係なく一般的な例について図6を参照しながら説明する。

【0052】

今、2つのクラスタC1、C2があるとする。クラスタC1として5個の文書D1、D2、D3、D4、D8が抽出され、クラスタC2には6個の文書D3、D4、D5、D6、D7、D8が抽出されたとする。

【0053】

ここで、2つのクラスタC1、C2に共通している文書は、文書D3、D4、D8である。この実施の形態では、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基に、それぞれのクラスタ間の関連性を判断してクラスタマージ処理を行う。

【0054】

具体的には、複数のクラスタととして、ある2つのクラスタに共通している文書数が2つのクラスタに存在する合計の文書数に対しどのくらいの割合かを計算し、その計算結果が予め定めたしきい値以上かどうかによってマージするか否かを定める。

【0055】

たとえば、この場合、2つのクラスタC1、C2に存在する文書数の合計は11個であり、両者に共通する文書数は3個である。これらから合計の文書数に占める共通する文書数の割合(%)を計算し、その結果からマージするか否かを決定する。この割合(%)を求める際、合計の文書数で共通する文書数を単純に割り算してそれに100を掛けて求めてもよいが、共通する文書数に任意に設定される係数を掛け算したものを合計の文書数で割り算してそれに100を掛けて求

めるようにしてもよい。

【0056】

一例として、クラスタC1に存在する文書数を $\alpha 1$ 、クラスタC2に存在する文書数を $\alpha 2$ とし、両者に共通する文書数を β とした場合、たとえば β に係数としてたとえば2を掛けて、 $2\beta / (\alpha 1 + \alpha 2) \times 100$ を計算し、その値(%)が予め設定されたしきい値TH(%)と比較して、上式による計算結果がしきい値TH以上であればマージするというようなことを行う。図6で示した例について考えれば、 2β は $2 \times 3 = 6$ 個、 $\alpha 1 + \alpha 2$ は $5 + 6 = 11$ 個であるので、この場合、約55%と求められる。ここで、しきい値THが仮に70%と設定されているとすれば、計算結果(55%)はしきい値TH(70%)より小さいので、クラスタC1とクラスタC2はマージしないとする。なお、係数は任意に設定されるもので、計算結果で得られる数値(%)がしきい値と比較し易いような値となるように適当に設定されるものであり、この場合は係数を2としたが、係数を1としても特に問題はない。

【0057】

ここで、図5で示した分類結果を例にして説明すれば、図5の場合、「用紙」のクラスタには文書D1, D4, D6, D7の4つの文書が存在し、「カセット」のクラスタには文書D1, D4, D7の3つの文書が存在する。そして、2つのクラスタに共通する文書は文書D1, D4, D7の3つの文書であり、これを合計の文書数に対する割合(%)で考える。

【0058】

これを前述した計算式によって計算する。図5の分類結果の場合、合計の文書数($\alpha 1 + \alpha 2$)は、 $4 + 3 = 7$ となり、共通の文書数は3で 2β は6となる。したがって、この場合、約86%という高い値が得られる。これは、設定されたしきい値(ここでは70%としている)よりも高いので、この「用紙」のクラスタと「カセット」のクラスタはマージして1つのクラスタとするということになる。

【0059】

同様に考えて、図5の「用紙」のクラスタと「増設」のクラスタとをマージす

るか否か、「カセット」のクラスタと「増設」のクラスタとをマージするか否かについて判断する。

【0060】

まず、「用紙」のクラスタと「増設」のクラスタについては、「用紙」のクラスタには文書D1, D4, D6, D7の4つの文書が存在し、「増設」のクラスタには文書D2, D3, D5, D7の4つの文書が存在する。そして、2つのクラスタに共通する文書は文書D7のみであり、これを上式を用いて計算すると、この場合、25%という結果が得られ、これは、しきい値(70%)よりも低い値であるので、この場合は、両者はマージしないとする。

【0061】

また、「カセット」のクラスタと「増設」のクラスタについては、「カセット」のクラスタには文書D1, D4, D7の3つの文書が存在し、「増設」のクラスタには文書D2, D3, D5, D7の4つの文書が存在する。そして、2つのクラスタに共通する文書は文書D7のみであり、これを上式を用いて計算すると、この場合、約28%という結果が得られ、これは、しきい値(70%)よりも低い値であるので、この場合は、両者はマージしないとする。

【0062】

このようにして、それぞれのクラスタに対し2つのクラスタごとにそれぞれマージするか否かを判断する。この図5の分類結果についてマージするか否かの処理を行ったあとの分類結果(マージ処理後の分類結果という)が図7である。図7によれば、「用紙」と「カセット」が「用紙+カセット」という1つのクラスタに分類され、そのクラスタに属する文書は文書D1, D4, D6, D7ということになる。また、「増設」についてはそのまま単独のクラスタを構成する。

【0063】

図7に示されるクラスタマージ処理後の分類結果において、たとえば、「用紙+カセット」で分類されたクラスタ(文書D1, D4, D6, D7が含まれる)について見れば、図2の文書内容からもわかるように、文書D1は用紙カセットについての内容であり、文書D4は用紙設定についての内容であり、文書D6は印刷された後の用紙の汚れた場合にはどのようにするかについての内容であり、

文書D7は用紙カセットの増設についての内容である。

【0064】

このように、これらの文書D1, D4, D6, D7はどれも用紙やカセットに関する内容であり、1つのクラスタとして分類されて何等问题のないものとなり、むしろ、「用紙+カセット」を1つのクラスタとした方がよい分類結果であるといえる。

【0065】

このように、始めにそれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理されて得られたそれぞれのクラスタに対し、2つずつのクラスタの組み合わせについてクラスタマージ処理を行うことによって、より適切なクラスタリングが行える。

【0066】

また、以上のようにして2つのクラスタごとに1回目のクラスタマージ処理が終了し、図7のようなクラスタマージ処理後の分類結果が得られると、今度は、そのクラスタマージ処理後の分類結果について、2回目のクラスタマージ処理を行う。つまり、図7の1回目のクラスタマージ処理後の結果で考えた場合、「用紙+カセット」のクラスタと「増設」のクラスタについてクラスタマージ処理を行う。この場合、「用紙+カセット」のクラスタと「増設」のクラスタについては、「用紙+カセット」のクラスタには文書D1, D4, D6, D7の4つの文書が存在し、「増設」のクラスタには文書D2, D3, D5, D7の4つの文書が存在する。そして、2つのクラスタに共通する文書は文書D7のみであり、これを合計の文書数に対する割合(%)で考えると、共通する文書数1に定数2を掛けたものを合計の文書数8で割り算し、それに100を掛けると、25%という結果が得られ、これは、しきい値(70%)よりも低い値であるので、この場合は、両者はマージしないとする。

【0067】

このようにして、2つのクラスタ間で1回目のクラスタマージ処理が終了した後、その1回目のクラスタマージ処理に新たな2つのクラスタ間で2回目のクラ

スタマージ処理を行い、その2回目のクラスタマージ処理が終了した後、その2回目のクラスタマージ処理後に新たな2つのクラスタ間で3回目のクラスタマージ処理を行うというクラスタマージ処理を順次行い、新たなクラスタが生成されなくなるまで（クラスタマージが起こらなくなるまで）その処理を繰り返す。

【0068】

また、これまでの説明は、2つのクラスタ間でクラスタマージ処理を行う例についてであるが、クラスタマージ処理は3つ以上のクラスタの組み合わせについても可能である。この場合、1回のクラスタマージ処理によって3つ以上のクラスタ間でクラスタマージ処理を行い、さらに、これによって幾つかのクラスタに分類された結果についてクラスタマージが起こらなくなるまで、順次、クラスタマージ処理を行うことも可能である。なお、3つ以上のクラスタについてクラスタマージするか否かを判断する場合、前述したように、それぞれのクラスタに存在する合計の文書数に対する共通の文書数の割合（％）で考えることができる。

【0069】

さらに、これまで説明した複数のクラスタ間でのクラスタマージ処理は、図5に示すような分類結果に基づき、それぞれのクラスタ間に共通する文書数が合計の文書数に占める割合を求め、それを設定されたしきい値との比較によって求めるようにしたが、このような方法によらず、それぞれのクラスタを特徴づける特徴要素が、元の文書においてどのような状態で用いられているかを調べることで、よってもクラスタマージ処理を行うことができる。これを実現するための文書分類装置の構成例を図8に示す。図8に示されるそれぞれの構成要素は図1と同じであり、同一部分には同一符号が付されているが、この場合、元の文書内容からクラスタマージするか否かを判断するため、クラスタマージ部2には、文書記憶部11の出力が与えられるようになっている。以下、これについて説明する。

【0070】

図5に示すような分類結果において、「用紙」のクラスタと「カセット」のクラスタをクラスタマージ処理する場合について説明する。「用紙」のクラスタには、文書D1、D4、D6、D7が含まれ、「カセット」のクラスタには、文書D1、D4、D7が含まれる。

【0071】

これら文書において、「用紙」と「カセット」がどのように用いられているかを調べる。文書D1においては、「用紙」と「カセット」が結びついた「用紙カセット」という用語が複数箇所出現し、文書D4には文書D1と同様に「用紙カセット」という用語が存在するとともに、「用紙」と「カセット」が近接した状態で用いられている。また、文書D7にも「用紙カセット」という用語や「用紙カセットユニット」という用語が存在する。また、文書D6には「カセット」という用語は存在しないが「用紙」という用語が複数出現する。

【0072】

これらのことから考えれば、特徴要素として抽出された「用紙」と「カセット」は、連続的に用いられたり近接して用いられたりすることの多い特徴要素であり、両者は関連性の高い特徴要素であることがわかる。このことから、少なくとも文書D1、D4、D7は関連性の高い文書であり、文書D6も全く関連性がないとは言えないので、この場合、「用紙」のクラスタと「カセット」のクラスタは「用紙+カセット」のクラスタとして1つにまとめても問題がないと判断できる。

【0073】

次に、「用紙」のクラスタと「増設」のクラスタをクラスタマージ処理する。「用紙」のクラスタには、文書D1、D4、D6、D7が含まれ、「増設」のクラスタには、文書D2、D3、D5、D7が含まれる。

【0074】

これら文書において、「用紙」と「カセット」がどのように用いられているかを調べる。文書D1、D2、D3、D4、D5、D6においては、「用紙」と「増設」が結びついて用いられた部分や、近接して用いられている部分はなく、文書D7のみにおいて「用紙カセット」と「増設」が近接した状態で用いられている程度である。

【0075】

したがって、これらのことから、特徴要素として抽出された「用紙」と「増設」は、連続的に用いられたり近接して用いられたりすることの多い特徴要素では

なく、両者はあまり関連性のある特徴要素であるとはいえないことがわかる。このことから、「用紙」のクラスタと「増設」のクラスタはマージしない方がよいということがわかる。

【0076】

また、「カセット」のクラスタと「増設」のクラスタをクラスタマージ処理すると、この場合も、「用紙」のクラスタと「増設」のクラスタにおけるクラスタマージ処理と同様に、「カセット」と「増設」が結びついて用いられた部分や、近接して用いられている部分は少ない。

【0077】

したがって、これらのことから、特徴要素として抽出された「カセット」と「増設」は、連続的に用いられたり近接して用いられたりすることの多い特徴要素ではなく、両者はあまり関連性のある特徴要素であるとはいえないことがわかる。このことから、「カセット」のクラスタと「増設」のクラスタはマージしない方がよいということがわかる。

【0078】

なお、このようなそれぞれのクラスタを特徴づける特徴要素が元の文書においてどのような状態で存在するかによってクラスタマージする処理においても、前述したように、それぞれのクラスタ間で1回目のクラスタマージ処理が終了した後、その1回目のクラスタマージ処理後に新たなクラスタ間で2回目のクラスタマージ処理を行い、その2回目のクラスタマージ処理が終了した後、その2回目のクラスタマージ処理後に新たなクラスタ間で3回目のクラスタマージ処理を行うというクラスタマージ処理を順次行い、新たなクラスタが生成されなくなるまで（クラスタマージが起こらなくなるまで）その処理を繰り返す。

【0079】

また、この場合も2つのクラスタ間でのクラスタマージ処理だけでなく、クラスタマージ処理は3つ以上のクラスタの組み合わせについても可能である。この場合、1回のクラスタマージ処理によって3つ以上のクラスタマージ処理を行い、さらに、これによって幾つかのクラスタに分類された結果についてクラスタマージが起こらなくなるまで、順次、クラスタマージ処理を行うことも可能である

【0080】

ところで、以上のようにしてクラスタマージ処理を行ったあと、クラスタマージされた後の結果をユーザに表示する際、どのような状況でクラスタマージを行ったのかを示す情報を付加情報としてユーザに提示することが好ましい。これは、クラスタマージ部2で行った処理内容を出力制御部31が受けてそれを表示部32に表示させるようにすることで行える。

【0081】

なお、本発明は以上説明した実施の形態に限定されるものではなく、本発明の要旨を逸脱しない範囲で種々変形実施可能となるものである。たとえば、前述の実施の形態では、図5に示すような分類結果を得るための特徴要素を各文書の表題部から得るようにして、表題部から得られた特徴要素に基づいたクラスタリングを行う例について説明したが、本発明においては、複数の文書を意味的に共通性のあるクラスタに分類し、その分類結果についてクラスタマージ処理を行うものである。複数の文書をクラスタリングする手法は、特に限定されるものではない。複数の文書をクラスタリングする手法としては、前述の実施の形態で説明した文書の表題部から得られた特徴要素に基づいてクラスタリングを行う例の他、たとえば、URLアドレス（たとえば、http://を取り除いた部分を使用する）、更新日時（単純な時間または最近1カ月以内の更新日時）、ファイルサイズ（webページ本文のバイトサイズなど）を用いてクラスタリングすることもできる。また、これらは、単独で用いてクラスタリングするようにしてもよく、幾つかを組み合わせてもよい。これらのどれを用いるかは、最初にメニューなどで選択項目を選ぶことで可能となる。また、選んだ項目が無い場合には、他の項目を代用する。たとえば、タイトルを選んだ場合、webページにタイトルが無い場合には、URLアドレスを代用する。

【0082】

そして、いずれかの方法によってクラスタリングされたのち、そのクラスタリング結果に対し、前述の実施の形態で説明したような処理、すなわち、それぞれのクラスタに含まれる文書の共通性を判断してそれぞれのクラスタ同志を統合す

るか否かを決めるという処理を施すことによってもクラスタマージを行うことができる。

【0083】

たとえば、URLによってクラスタリングする場合について説明すれば、あるURL（これをURL1とする）のクラスタと、あるURL（これをURL2とする）のクラスタに分類されたとし、URL1のクラスタには文書D1, D2, D3, D4が存在し、URL2のクラスタには文書D2, D3, D4, D5が存在したとする。この場合、これら2つのクラスタには、共通する文書として文書D2, D3, D4が含まれることになり、この共通する文書数と合計の文書数との関係から、URL1のクラスタとURL2のクラスタを統合するか否かを決める。

【0084】

また、クラスタマージするか否かの判断は、前述の実施の形態では、対象となるクラスタに含まれる合計の文書数で共通の文書数を割って得られる割合（％）で表し、その値が予め設定されたしきい値（％）と比較することによって行ったが、これに限られるものではなく、たとえば、共通する文書の個数を数え、その個数とそれぞれのクラスタに含まれる文書数との関係からマージするかしないかを決めるようにすることも可能である。

【0085】

また、前述の実施の形態では、文書D1, D2, …, D7は、それぞれが独立した文書であって、それぞれ独立した文書を分類する場合について説明したが、ある1つの文書を幾つかのコンテンツに分けて、それぞれのコンテンツ（ここでいうコンテンツとは文書の中の意味的なまとまりを指す）を分類する場合にも適用できる。ここで抽出されるコンテンツは、各表題部ごとに切り分けられて得られる文書の中の意味的なまとまりであるとする。

【0086】

たとえば、図2で示した文書D1, D2, …, D7が集まって1つの文書が構成されていると仮定すれば、文書D1, D2, …, D7をそれぞれコンテンツとみなすことができる。これらをコンテンツとすれば、それぞれのコンテ

ンツは、表題部 T 1, T 2, . . . , T 7 と本文 A 1, A 2, . . . , A 7 から構成されたものとなる。

【0087】

このように、1つの文書を複数のコンテンツに分けて考えた場合、本発明はそれぞれのコンテンツをクラスタリングし、そのクラスタリング結果をクラスターマージする場合にも同様に適応できる。

【0088】

さらに、本発明で用いられるクラスタリング対象文書は、たとえば、汎用の検索サービスで検索された複数の文書をクラスタリング対象文書として考えることもできる。この場合、検索された多数の文書に対してクラスタリング処理を行い、そのクラスタリングされた結果についてクラスターマージ処理を行う。

【0089】

また、以上説明した本発明の文書分類処理を行う処理プログラムは、フロッピーディスク、光ディスク、ハードディスクなどの記録媒体に記録しておくことができ、本発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてもよい。

【0090】

【発明の効果】

以上説明したように本発明によれば、それぞれの文書を複数のクラスタに分類したのちに、その複数のクラスタ間で各々のクラスタに含まれる文書の内容に基づいて各々のクラスタ間の関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスターマージ処理を行うようにしている。これによって、最初のクラスタ処理によって、多数のクラスタが生成されたとしても、それぞれのクラスタ間での関連性を判断して複数のクラスタを統合することができるので、簡潔化された見易いクラスタリング結果をユーザに提示することができ、ユーザは自分の欲しい情報を効率よく探すことができるようになる。

【0091】

また、クラスタ間の関連性の判断は、クラスターマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、各々のクラスタに共通して含まれる文書数

を基にして行うので、簡単で的確なクラスタマージ処理を行うことができる。

【0092】

また、クラスタ間の関連性の判断を行うための他の方法として、特徴要素がクラスタマージ処理対象となるクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージ処理を行うようにしてもよく、これによれば、実際の文書内容に基づいたクラスタ同志の関連性の判断が行えるので、適切なクラスタマージ結果を得ることができる。

【0093】

そして、クラスタマージ処理は、少なくとも2つのクラスタの組み合わせで行い、さらに、所定の数のクラスタの組み合わせでのクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すことによって、最終的には、より簡潔的に整理されたクラスタリング結果とすることができる。

【0094】

また、このようなクラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することにより、ユーザはどのような状況でクラスタマージ処理がなされたかを知ることができるので、クラスタマージ処理後の結果から自分の欲しい情報を探す際に、その付加状況を参考にして探すことができる。

【0095】

このように本発明によれば、複数の文書を分類処理することによって生成された多数のクラスタに対し、クラスタマージ処理を施すことにより関連性の高いクラスタ同志を統合してまとめることができるので、たとえば、検索サービスなどで検索された多数の文書に対し、本発明を適用することにより、検索要求を出したユーザに対し、検索結果を簡潔化した見易いクラスタリング結果として提示することができる。これによって、ユーザは自分の欲しい情報を効率よく探すことができ、従来にはない検索サービスが実現できる。

【図面の簡単な説明】

【図 1】

本発明の文書分類装置の実施の形態を説明するブロック図である。

【図 2】

本発明の実施の形態を説明するための複数の文書例を示す図である。

【図 3】

本発明が行う文書分類処理の処理手順を概略的に説明するフローチャートである。

【図 4】

特徴要素とそれぞれの文書との関係を示す特徴テーブル内容の一例を示す図である。

【図 5】

図 4 に示す特徴テーブルに基づいて文書を分類した分類結果を示す図である。

【図 6】

2 つのクラスタ間でのクラスタマージ処理を説明する図であり、それぞれのクラスタに含まれる文書例を示す図である。

【図 7】

図 5 の分類結果についてクラスタマージ処理した結果を示す図である。

【図 8】

特徴要素が元の文書にどのように出現するかによってクラスタマージを行う場合の文書分類装置のブロック図である。

【符号の説明】

- 1 クラスタリング部
- 2 クラスタマージ部
- 3 分類結果出力部
- 1 1 文書記憶部
- 1 2 文解析部
- 1 3 特徴要素抽出部
- 1 4 特徴テーブル作成部

1 5 文書分類部

1 6 分類結果記憶部

3 1 出力制御部

3 2 表示部

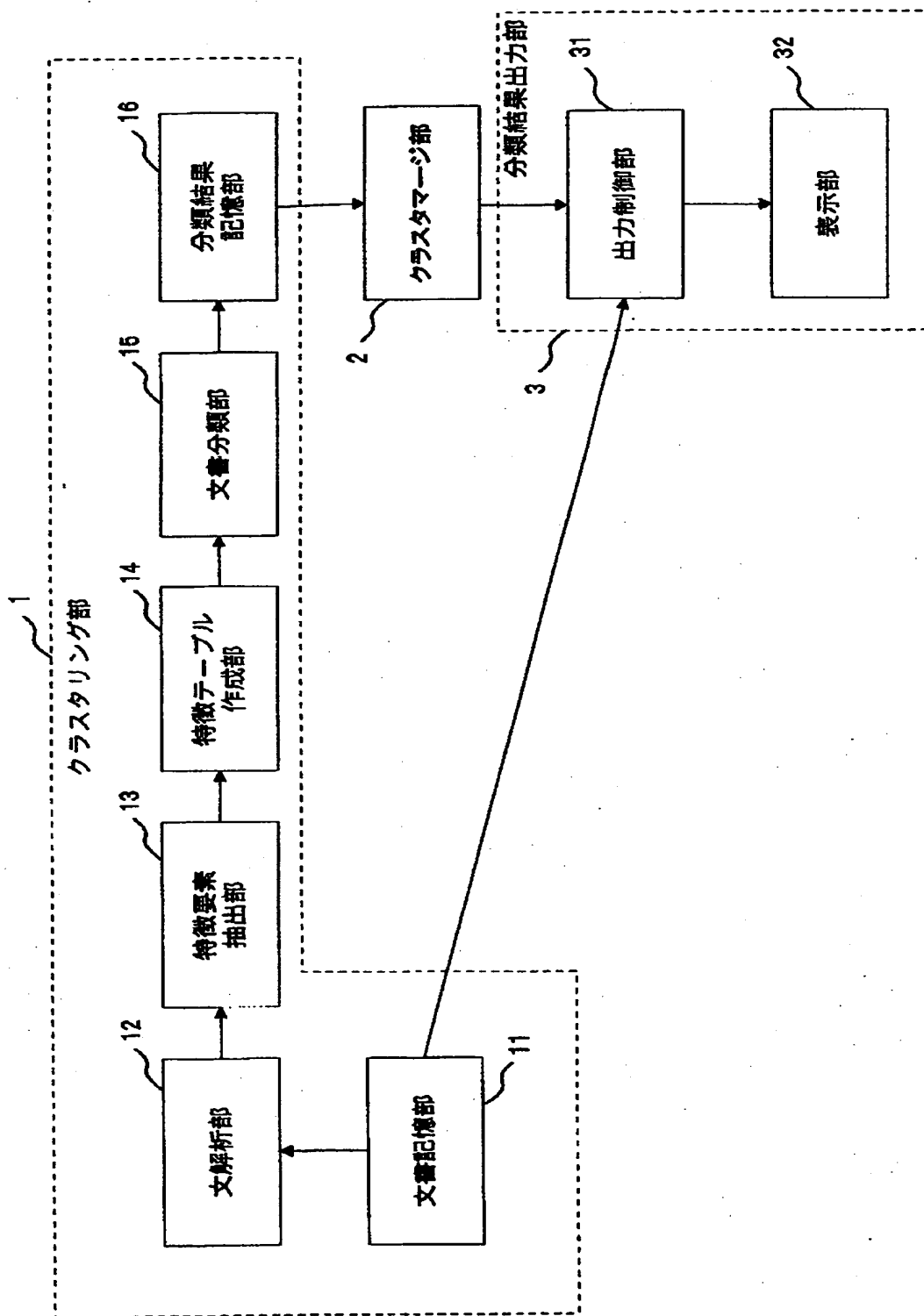
A 1, A 2, . . . , A 7 本文

D 1, D 2, . . . , D 7 文書

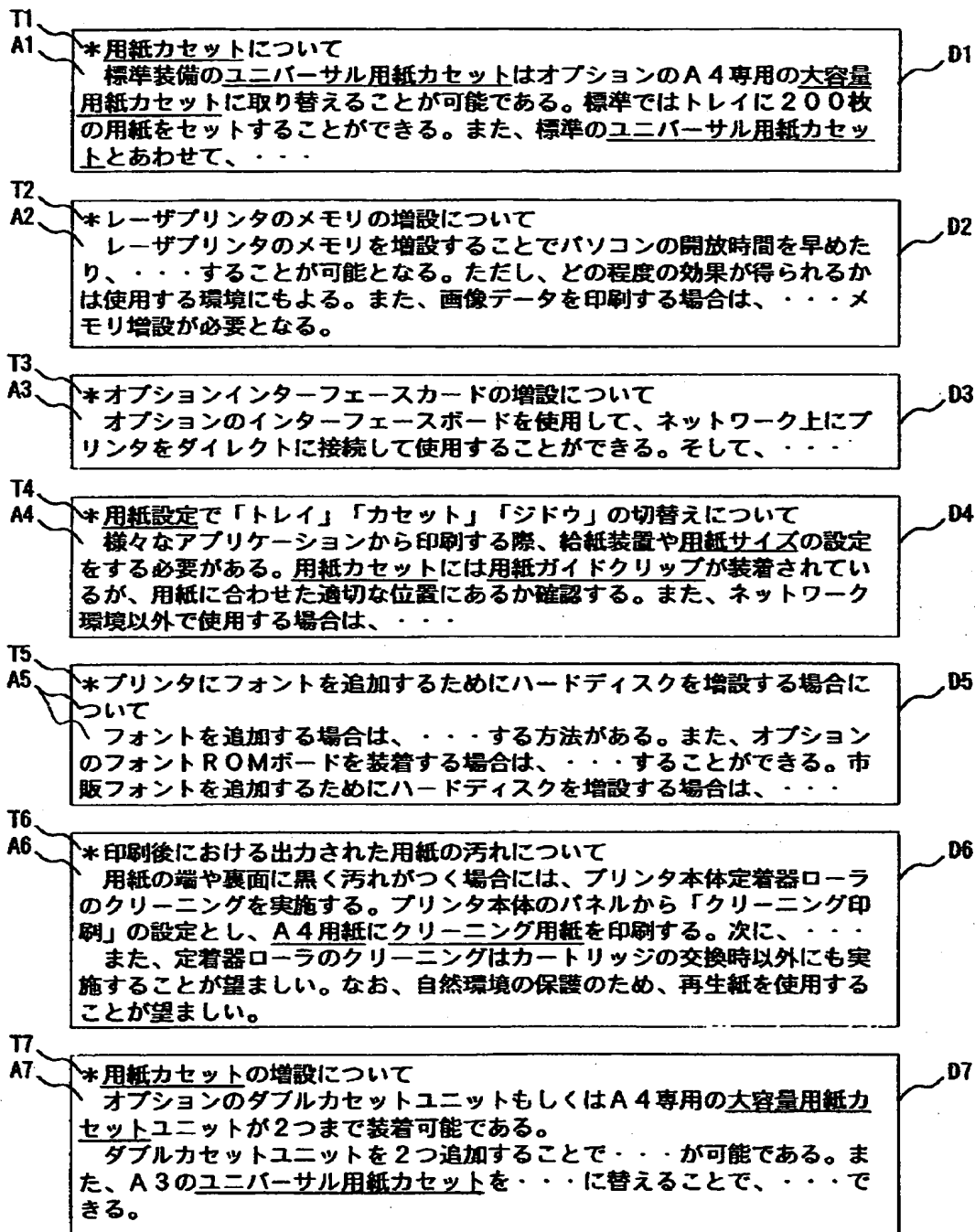
T 1, T 2, . . . , T 7 表題部

【書類名】 図面

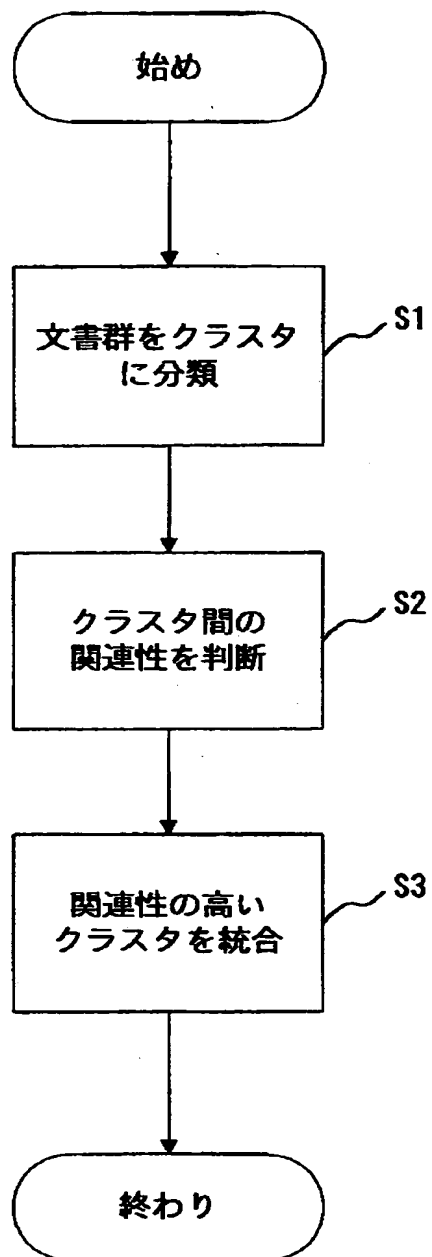
【図 1】



【図 2】



【図3】



【図 4】

特徴要素	文書 D 1	文書 D 2	文書 D 3	文書 D 4	文書 D 5	文書 D 6	文書 D 7
用紙	1			1		1	1
カセット	1			1			1
増設		1	1		1		1

【図 5】

特徴要素	クラスタ
用紙	D 1, D 4, D 6, D 7
カセット	D 1, D 4, D 7
増設	D 2, D 3, D 5, D 7

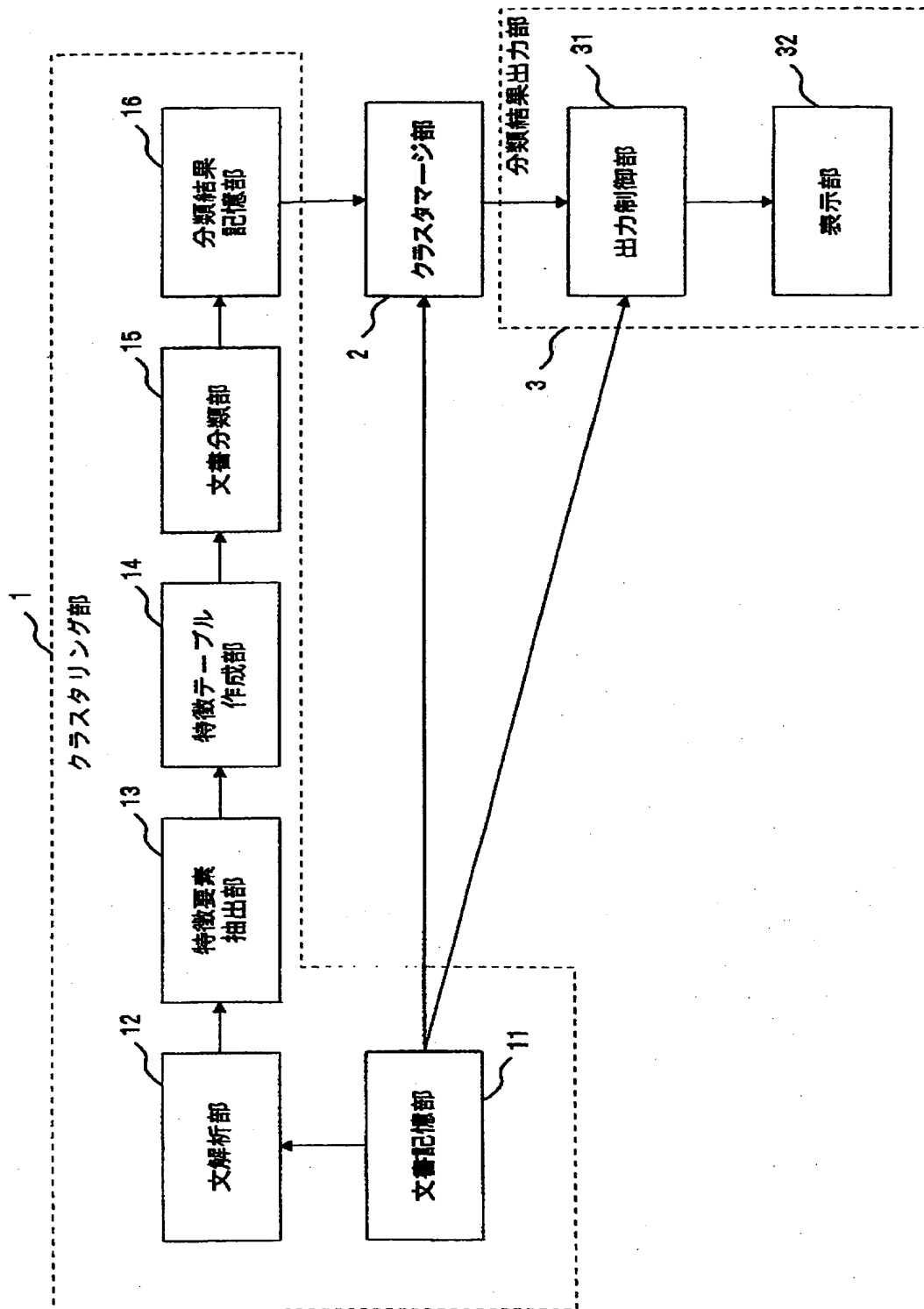
【図 6】

クラスタ C1	D 1, D 2, D 3, D 4, D 8
クラスタ C2	D 3, D 4, D 5, D 6, D 7, D 8

【図 7】

特徴要素	クラスタ
用紙+カセット	D 1, D 4, D 6, D 7
増設	D 2, D 3, D 5, D 7

【図 8】



【書類名】 要約書

【要約】

【課題】 多数の文書をそれぞれの文書に存在する特徴要素に基づいてクラスタに分類する場合、多数のクラスタを整理して出力する。

【解決手段】 複数の文書をそれぞれ解析して表題部を検出する文解析部 12 と、文解析部 12 で検出されたそれぞれの処理対象文書の表題部から特徴要素を抽出する特徴要素抽出部 13 と、表題部から抽出された特徴要素とその特徴要素を含む処理対象文書との関係を示す特徴テーブルを作成する特徴テーブル作成手段 14 と、作成された特徴テーブルの内容を参照して前記処理対象文書を意味的に共通性を有する複数のクラスタに分類する文書分類部 15 と、文書分類部 15 により分類されたクラスタを記憶する分類結果記憶部 16 と、分類結果記憶部 16 に記憶されたクラスタをクラスタマージ処理するクラスタマージ部 2 と、そのクラスタマージ処理結果を表示部 32 に出力する出力制御部 31 とを有した構成とする。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000002369]

1. 変更年月日 1990年 8月20日
[変更理由] 新規登録
住 所 東京都新宿区西新宿2丁目4番1号
氏 名 セイコーエプソン株式会社